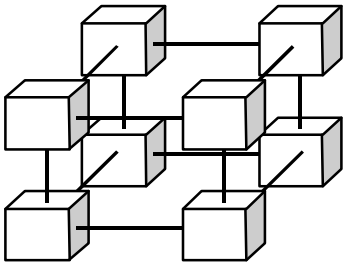
# BlueGene/L Hardware

- BG/L networks
- Compute ASIC
- Partitioning

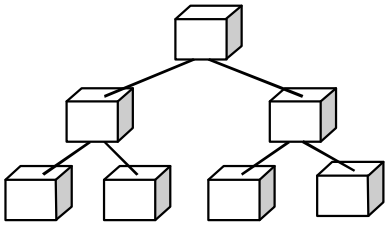
Al Gara
IBM Research

# BlueGene/L - Five Independent Networks



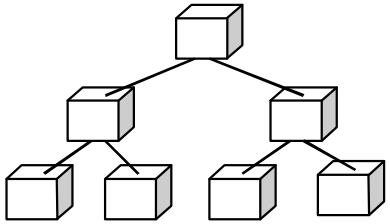
## 3 Dimensional Torus

- Point-to-point



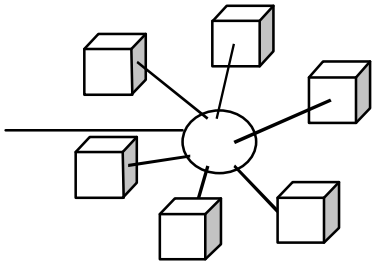
## Global Tree

- Global Operations



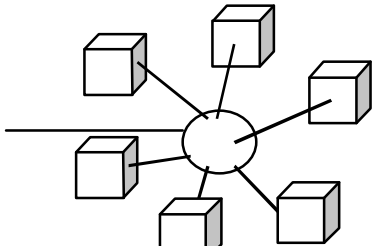
## Global Barriers and Interrupts

- Low Latency Barriers and Interrupts



## Gbit Ethernet

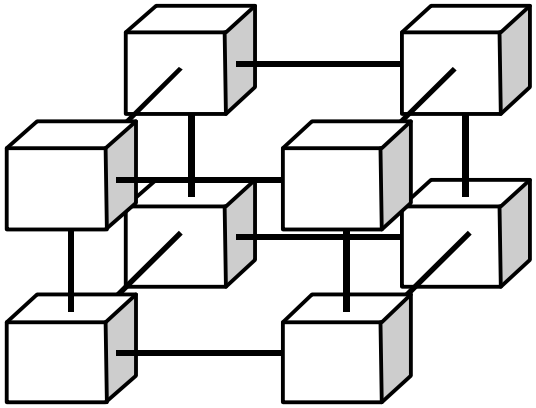
- File I/O and Host Interface

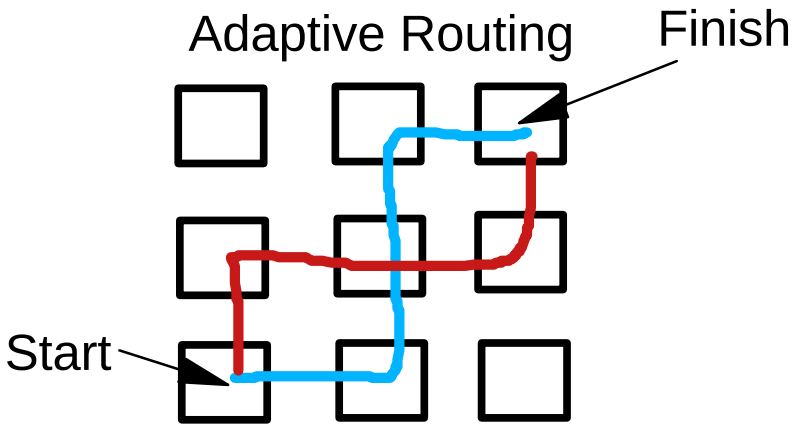


## Control Network

- Boot, Monitoring and Diagnostics

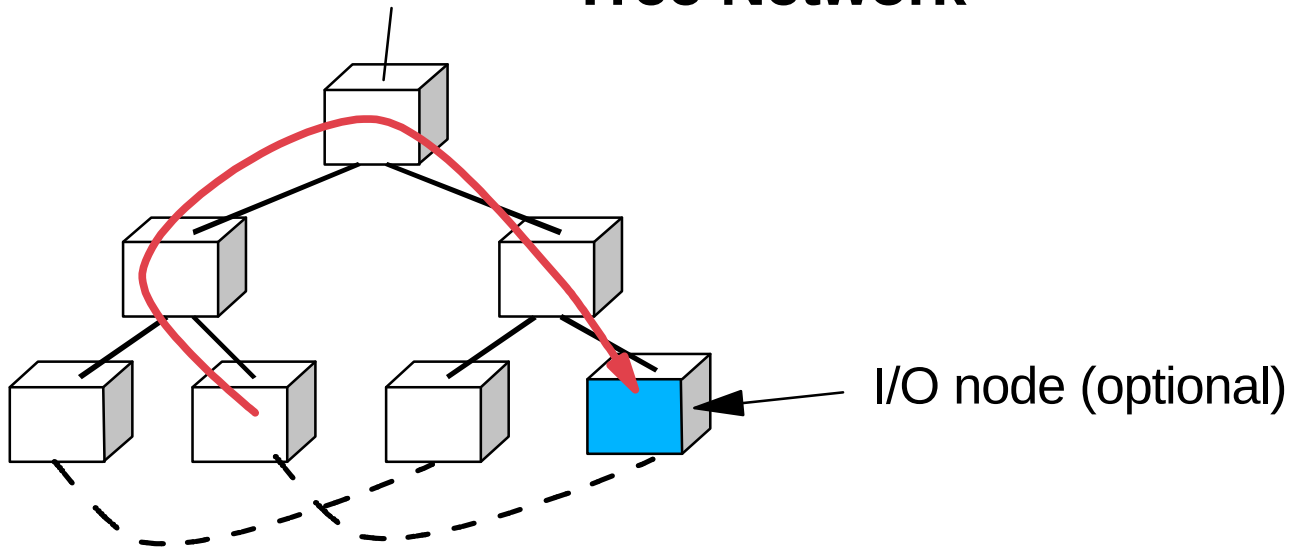
# Three-dimensional Torus Network





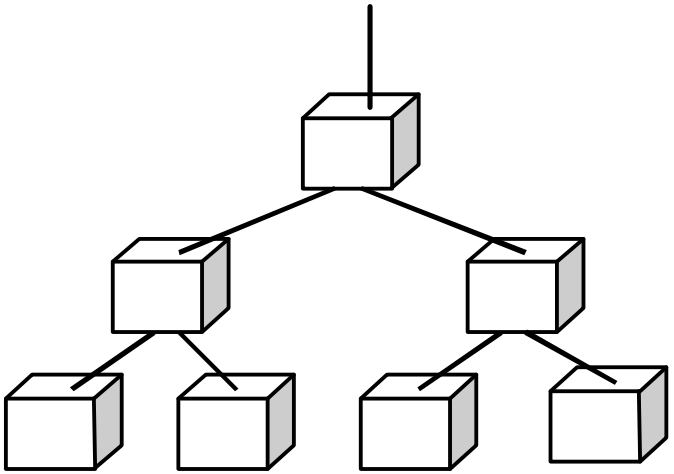

- 32x32x64 connectivity
- Backbone for one-to-one and one-to-some communications
- 1.4 Gb/s bi-directional bandwidth in all 6 directions (Total 2.1 GB/s/node)
- 64k * 6 * 1.4Gb/s = 68 TB/s total torus bandwidth
- 4 * 32 *32 * 1.4Gb/s = 5.6 Tb/s Bisectional Bandwidth
- Worst case hardware latency through node ~ 69nsec
- Virtual cut-through routing with multipacket buffering on collision
  - Minimal
  - Adaptive
  - Deadlock Free
- Class Routing Capability (Deadlock-free Hardware Multicast)
  - Packets can be deposited along route to specified destination.
  - Allows for efficient one to many in some instances
- Active messages allows for fast transposes as required in FFTs.
- Independent Channels and Control for each Node Processor

# Tree Network




- High Bandwidth one-to-all
  - 2.8Gb/s to all 64k nodes
  - 68TB/s aggregate bandwidth
- Arithmetic operations implemented in tree
  - Integer/ Floating Point Maximum/Minimum
  - Integer addition/subtract
- Latency of tree less than 2.5usec to top, additional 2.5usec to broadcast to all
- Global sum over 64k in less than 2.5 usec (to top of tree)
- Used for disk/host funnel in/out of I/O nodes.
- Minimal impact on cabling
- Partitioned with Torus boundaries
- Flexible local routing table
- Good scaleability
- Fault Tolerant
- Used as Point-to-point for File I/O

# Fast Barriers

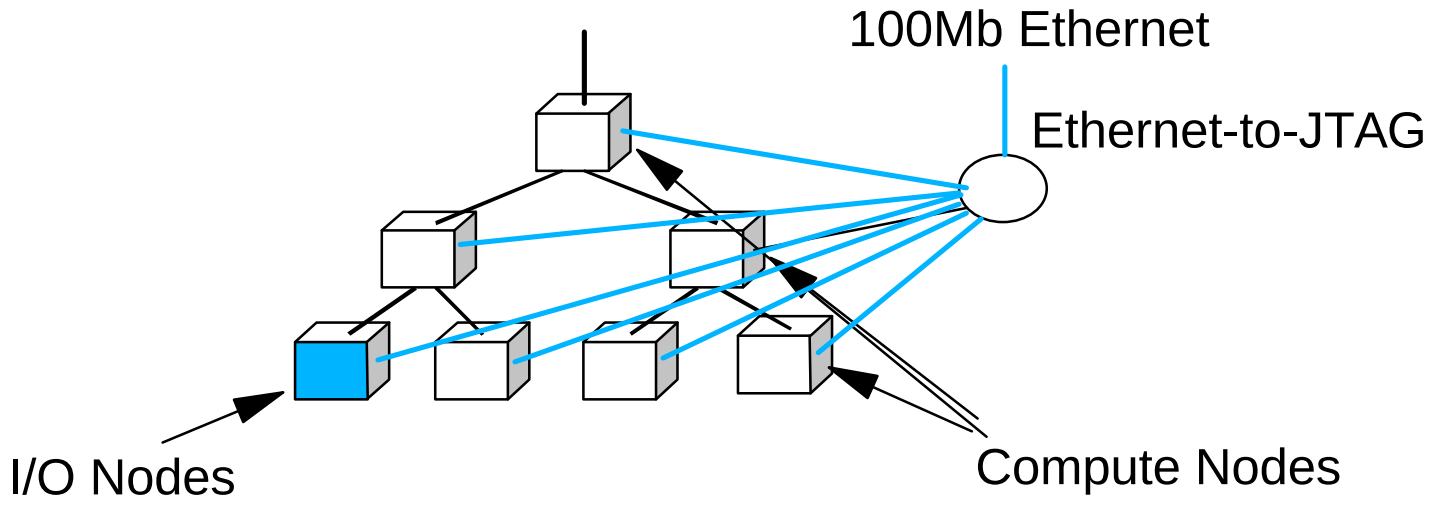


- Four Independent Barrier or Interrupt Channels
  - Independently Configurable as "or" or "and"
- Asynchronous Propagation
  - Halt operation quickly (current estimate is 1.3usec worst case round trip)
    > 3/4 of this delay is time-of-flight.
- Sticky bit operation
  - Allows global barriers with a single channel.
- User Space Accessible
  - System selectable
- Partitions along same boundaries as Tree, and Torus
  - Each user partition contains it's own set of barrier/ interrupt signals

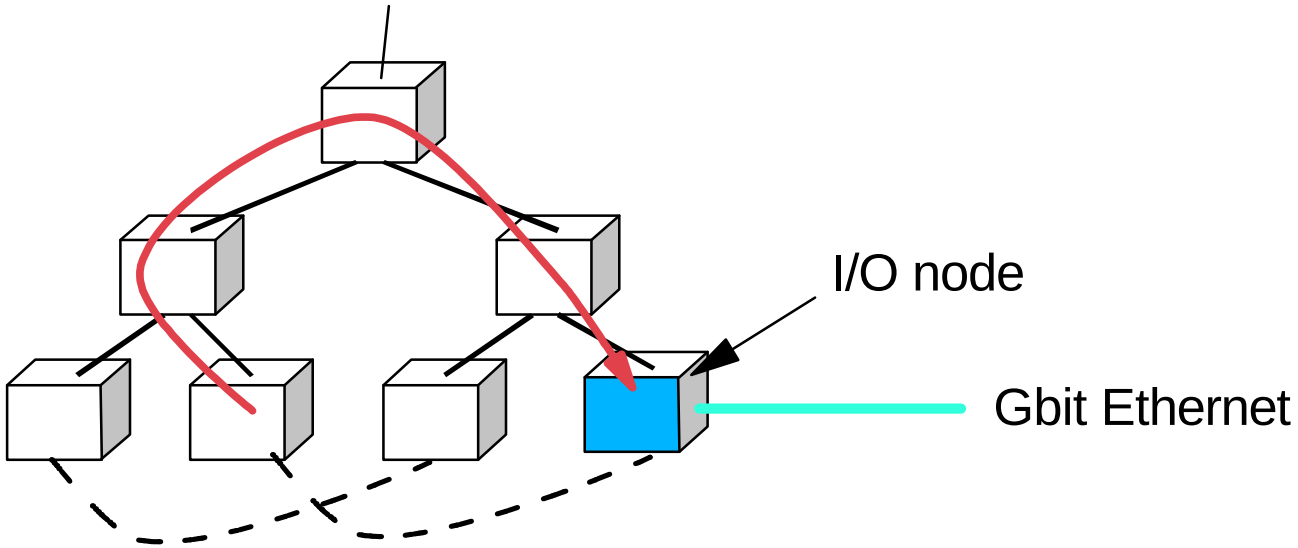
# Control Network

## JTAG interface to 100Mb Ethernet

- direct access to all nodes.
- boot, system debug availability.
- runtime noninvasive RAS support.
- non-invasive access to performance counters
- Direct access to shared SRAM in every node



# Ethernet Disk/Host I/O System




## Gb Ethernet on all I/O nodes

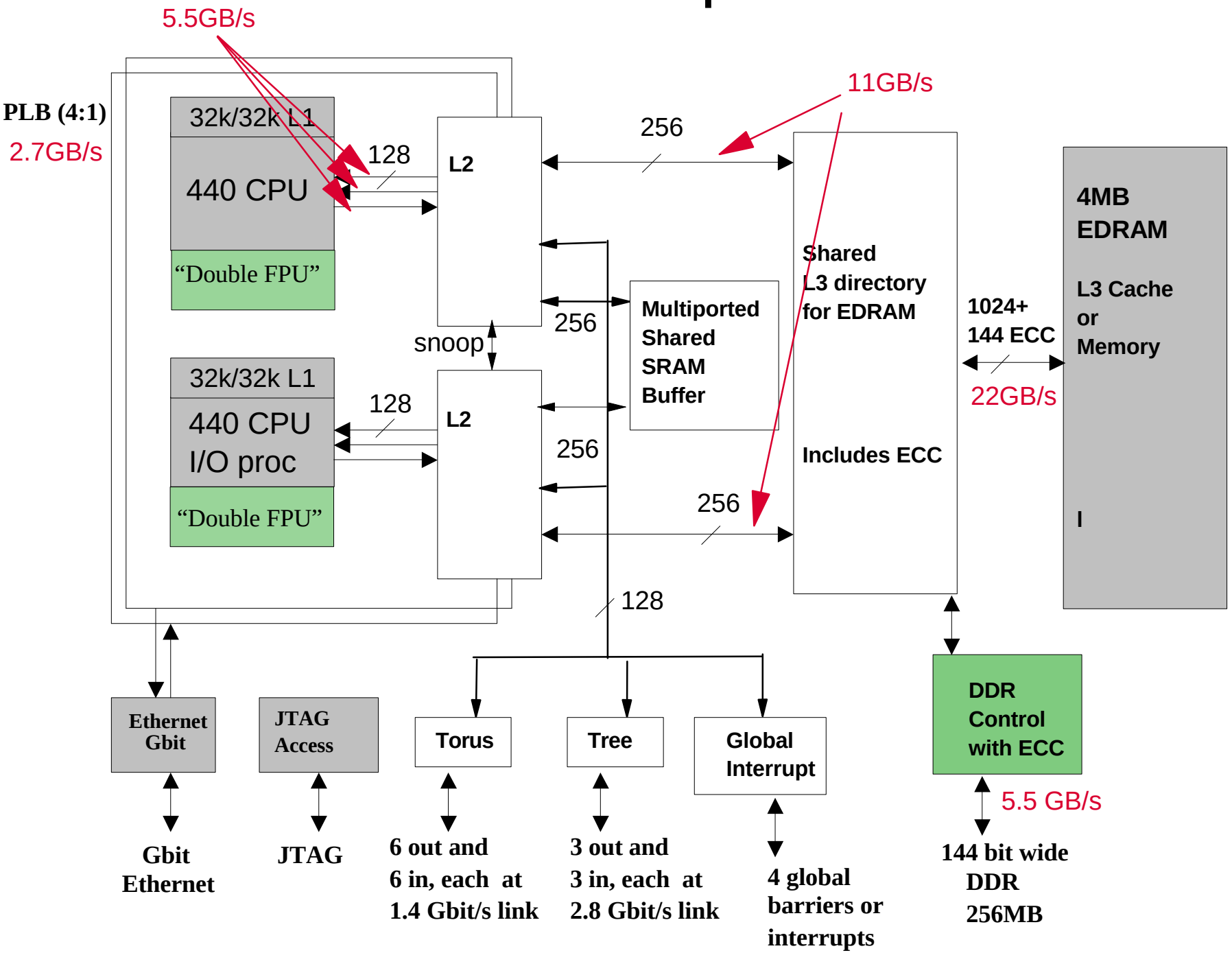
- **Gbit Ethernet Integrated in all node ASICs but only used on I/O nodes.**
- **Funnel via global tree.**
- **I/O nodes use same ASIC but are dedicated to I/O Tasks.**
- **I/O nodes can utilize larger memory.**

## Dedicated DMA controller for transfer to/from Memory

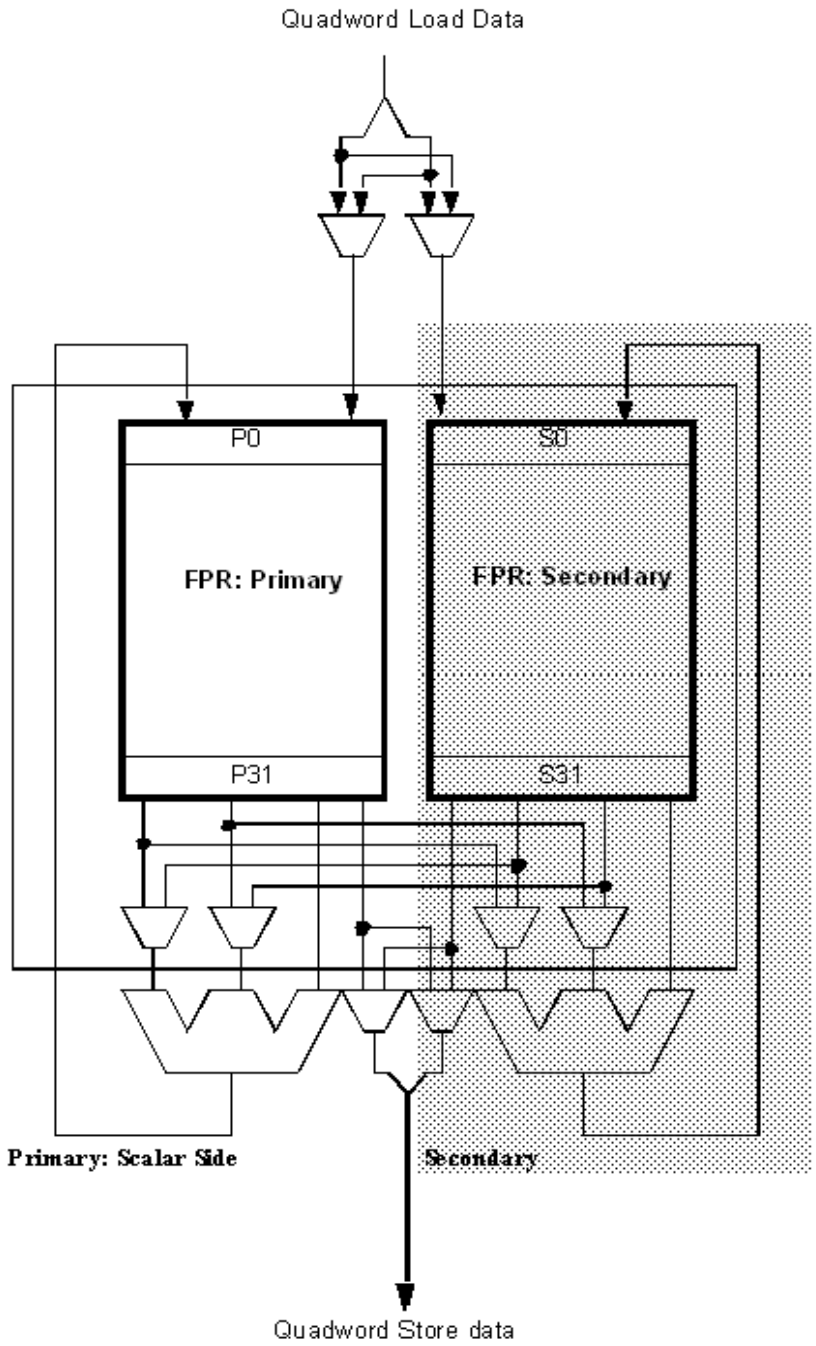
## Configurable ratio of Compute to I/O nodes

- **I/O nodes are leaves on the tree network**

# BlueGene/L Compute ASIC



# Floating Point Unit



# Level 2 Cache

## Independent L2 Caches

- 16 128B L2 Cache lines (fully associative)
- Low Latency Interface to Processor Cores
  - 1-2 processor cycles additional latency

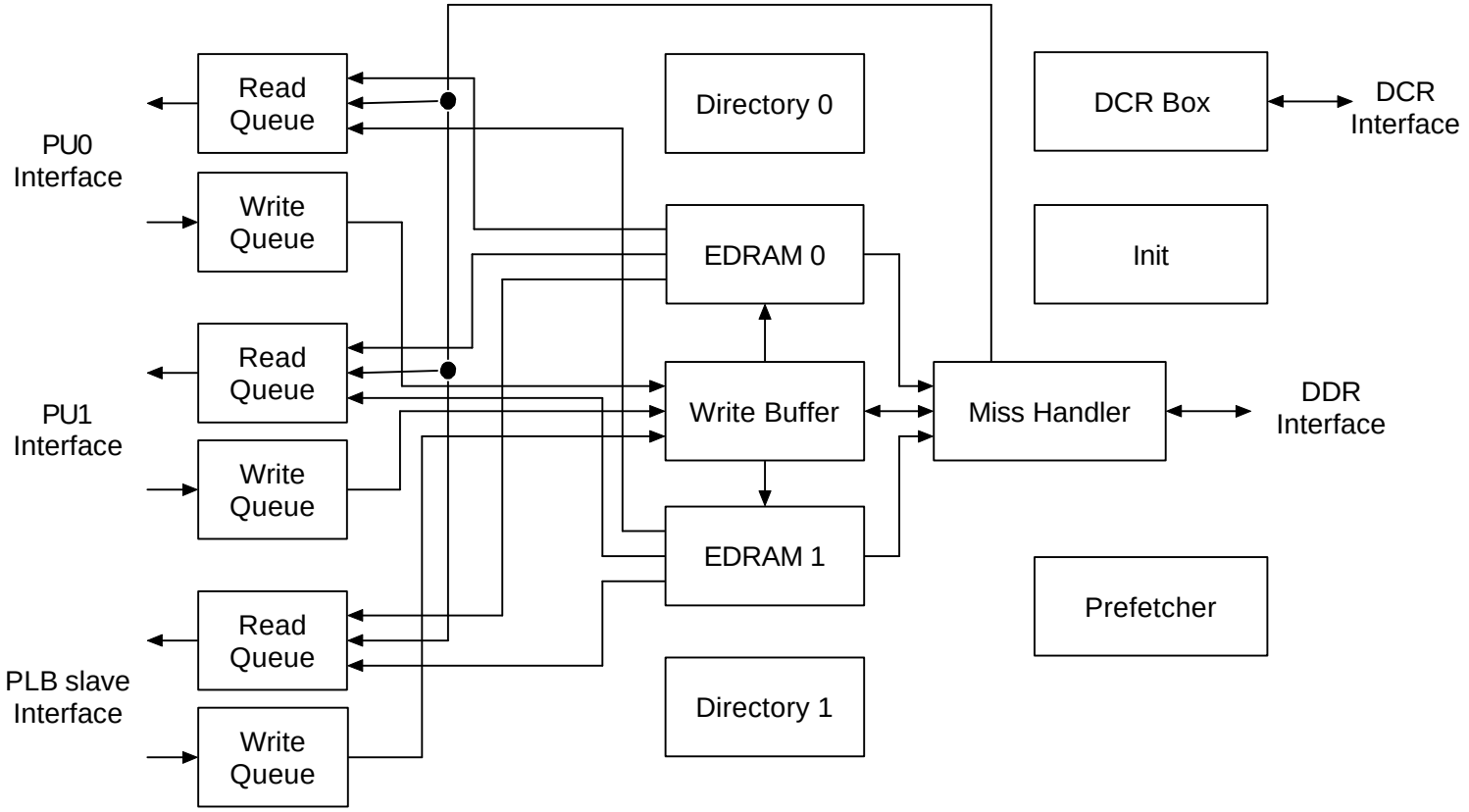
## Prefetching

- Configurable prefetching modes
  - None
  - Always
  - Confirmed

## Coherency

- Memory system is coherent outside the L1 caches

# Level 3 Cache




## 4 MB eDRAM based L3 cache

- Organized as two 8-way set associative banks
- Simultaneous access to both banks
- 22 GB/s peak bandwidth

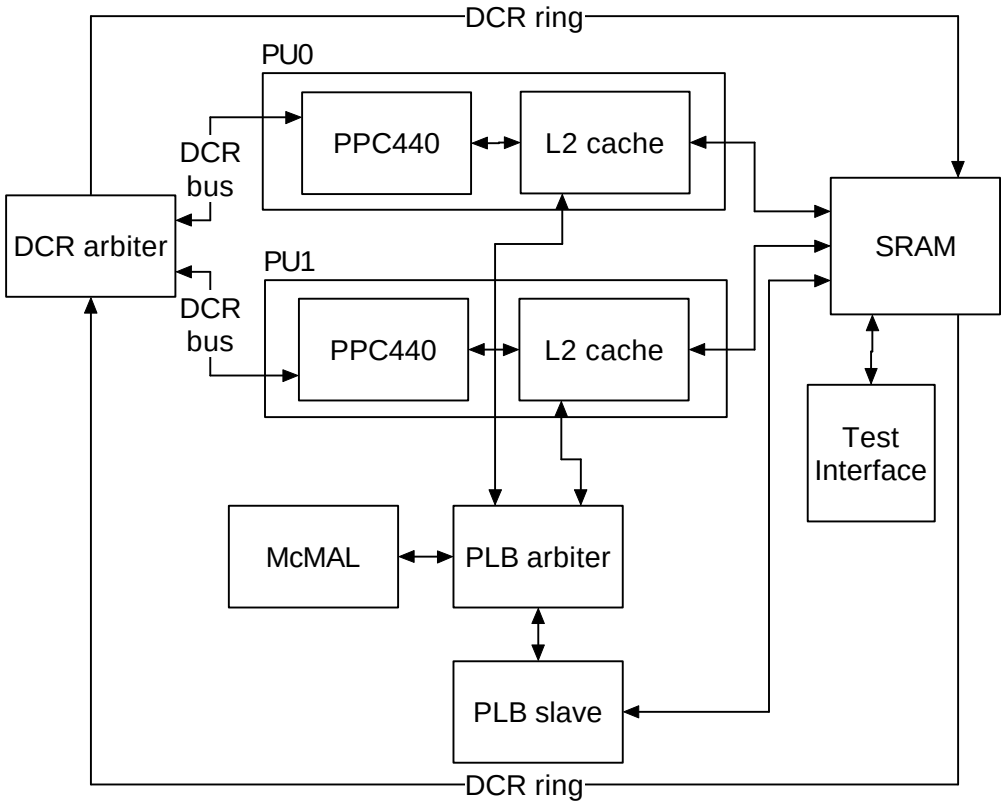
## Prefetching

- Programmable Prefetch Depth

## Partitionability

- Flexible configuration of memory/cache mapped in 256KB increments

# Shared SRAM




**Low Latency (~ 4 cycles) Access to small SRAM**

- **Simultaneous Access from both processor units**
- **Used in conjunction with Lock Box for messaging**

# Shared Hardware Locks / Blind Device

**Locks provide atomic test and set**

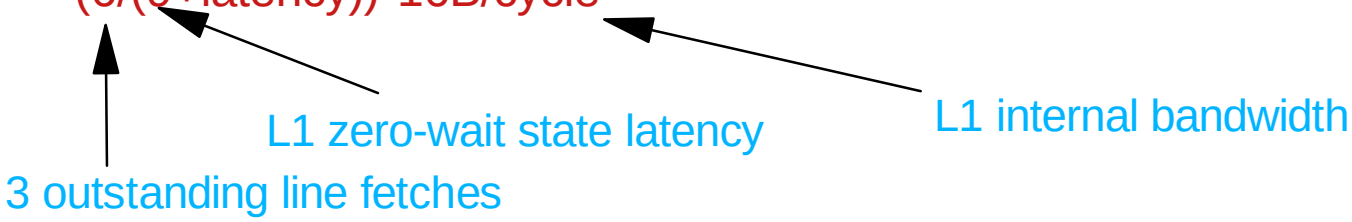
- **Accessible with a single load operation**
- **256 independent single bit locks**
- **Load returns the status of the lock**

- **Blind device provides a dummy read address**
  - **Assists in L1 cache control (proper reads can clear cache)**

# Memory System

atency and Bandwidth estimates

- Quadload
- random assumes (1/2 cache line)
- Bandwidth for Random Access (6/(9+latency))*16B/cycle




| | latency | Sustained Bandwidth Random Access | Sustained Bandwidth Sequential Access |
|---|---|---|---|
| L1 | 3 | 16.0B/cycle | 16.0B/cycle |
| L2 | 11 | 2.7B/cycle | 5.3B/cycle |
| SRAM | 15 | 2.0B/cycle | 5.3B/cycle |
| L3 (eDRAM page hit) | 23 | 1.5B/cycle | 5.3B/cycle |
| L3 (eDRAM page miss) | 31 | 1.2B/cycle | (NA) |
| External DRAM (single processor) | 75 | 0.57 B/cycle | 5.3B/cycle |
| External DRAM(dual processor) | 75 | 0.57 B/cycle | 4.0B/cycle |
| | | | |
| | | | |

# Partitioning for Accessibility and Reliability

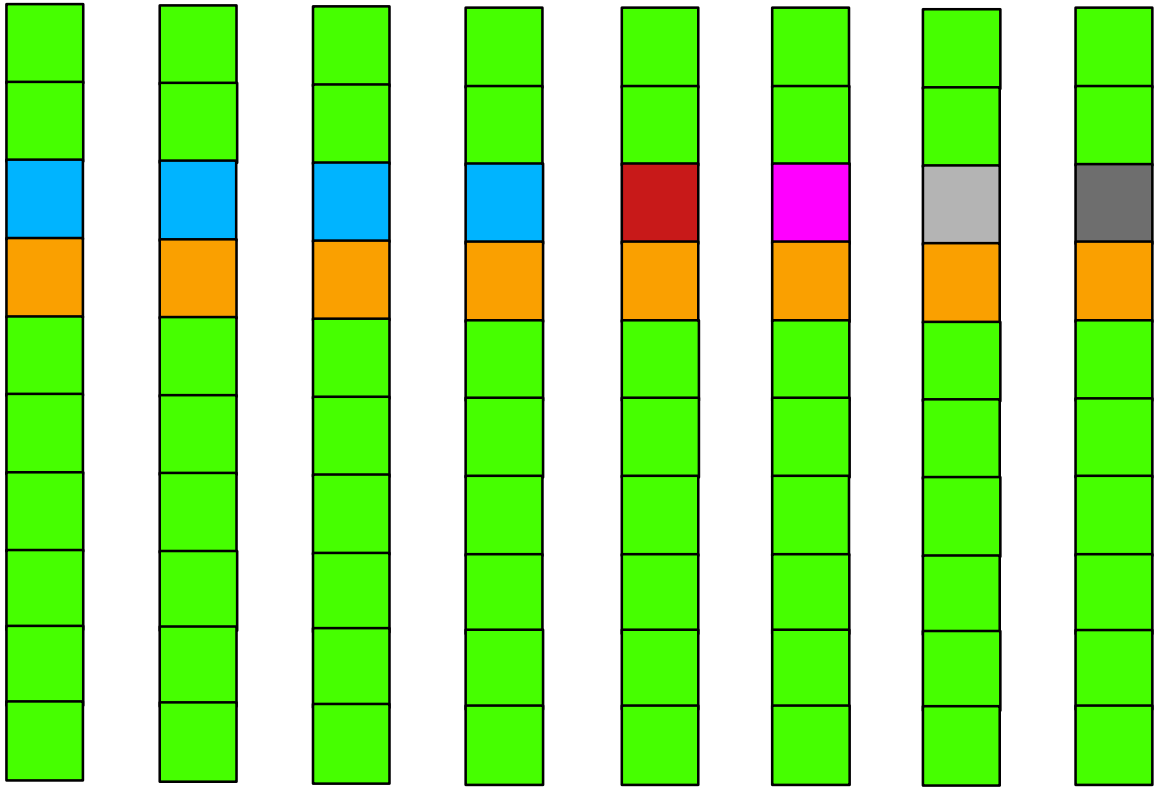
**Partitioning:**

- Multiple User machine accomplished through space partitioning.
- All networks (torus, global tree and Ethernet I/O ) partition together. Users are protected from other partitions through hardware.
- Performance is not affected by partitioning. (some longer cables is the only effect)

**Reliability:**

- Partitioning allows for sectors to be swapped in for known bad sectors.
- Packaging allows for simple non-intrusive access to bad node cards.

Top View of BlueGene/L



**Possible rack partitioning:**

**1@ 64k node system**
**+ 1@ 8k system**
**+ 1@ 4k system**
**+ 4 @1k systems**